ARI Research Note 91-03

Test Analysis Program Evaluation: Item Statistics as Feedback to Test Developers

Peter J. Legree

U.S. Army Research Institute

Field Unit at Fort Gordon, Georgia
Michael G. Sanders, Chief

Training Research Laboratory
Jack H. Hiller, Director

October 1990

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Technical review by

William J. York, Jr.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

REPORT SECURITY CLASSIFICATION nclassified	1b. RESTRICTIVE MARKINGS --
SECURITY CLASSIFICATION AUTHORITY -	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
DECLASSIFICATION/DOWNGRADING SCHEDULE -	
PERFORMING ORGANIZATION REPORT NUMBER(S) ARI Research Note 91-03	5. MONITORING ORGANIZATION REPORT NUMBER(S) --
NAME OF PERFORMING ORGANIZATION U.S. Army Research Institute Fort Gordon Field Unit 6b. OFFICE SYMBOL (If applicable) PERI-IG	7a. NAME OF MONITORING ORGANIZATION --
ADDRESS (City, State, and ZIP Code) Attn: PERI-IG (Bldg. 41203) Fort Gordon, GA 30905-5230	7b. ADDRESS (City, State, and ZIP Code) --
NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI-I	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER --
ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 63007A; PROJECT NO. 795; TASK NO. 3303; WORK UNIT ACCESSION NO. H01

TITLE (Include Security Classification)

Test Analysis Program Evaluation: Item Statistics as Feedback to Test Developers

PERSONAL AUTHOR(S)
Legree, Peter J.

13a. TYPE OF REPORT Final	13b. TIME COVERED FROM 89/02 TO 90/03	14. DATE OF REPORT (Year, Month, Day) 1990, October	15. PAGE COUNT 16

SUPPLEMENTARY NOTATION
--

COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Test analysis; Item analysis; Reliability; Systems Approach to Training (SAT); Item consistency; Validity; Item difficulty

ABSTRACT (Continue on reverse if necessary and identify by block number)

The test analysis program was evaluated to determine the feasibility of using a personal computer to provide course developers with item statistics. This project was undertaken because of Signal School concern that course tests do not accurately measure student school performance. The evaluation focused on the usefulness of providing item statistics to course test developers and demonstrated that many of the tests contain poorly written items. The evaluation indicates that a computerized test analysis program can be used to identify questionable test items and help ensure Signal School tests are adequate to validate lessons and courses.

DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
a. NAME OF RESPONSIBLE INDIVIDUAL Peter J. Legree	22b TELEPHONE (Include Area Code) (404) 791-5523/5524 22c. OFFICE SYMBOL PERI-IG

Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

TEST ANALYSIS PROGRAM EVALUATION: ITEM STATISTICS AS FEEDBACK TO TEST DEVELOPERS

CONTENTS

	Page
INTRODUCTION AND BACKGROUND.	1
Statement of Problem.	1
Criterion Referenced Testing at the Signal School.	1
Description of SQT Test Item Development and Standards.	2
Test Analysis Program Recommendation.	4
METHOD	4
RESULTS AND DISCUSSION	5
Feedback to the Course Developers	5
Item Statistics Summary	5
CONCLUSIONS.	9
REFERENCES	11
APPENDIX A. RELATIONSHIP BETWEEN TEST RELIABILITY AND TRAINING EVALUATION	A-1

LIST OF TABLES

Table 1. Summary statistics for the 29E and 31M tests.	6
2. Proportion of test items not meeting ATSC standards	7

TEST ANALYSIS PROGRAM EVALUATION: ITEM STATISTICS AS FEEDBACK TO TEST DEVELOPERS

Introduction and Background

This project was undertaken to address the concern that performance on Signal course tests does not accurately measure student school performance. The project goal was to assess one method that could be used to improve Signal School test quality: a computerized item analysis program that identifies questionable test items.

The Signal School is concerned with test accuracy because the quality of the school's lessons is partially dependent on the quality of the course tests since changes in the lessons must be validated using actual students (Tradoc Regulation 350-7, 1988). This policy applies to both minor and major lesson modifications and revisions, as well as to the implementation of new training technologies and approaches to teach lessons. Course test quality is important because practical constraints usually limit the collection of the validation data to student performance on course tests. The Signal School also recognizes that the accuracy of the course test scores limits the quality of student-related managerial decisions and is important to the maintenance and development of student motivation.

Statement of Problem

The Deputy Assistant Commandant at the Signal School requested assistance from the Army Research Institute to implement procedures to improve the accuracy of Signal School tests. This research project focused on methods that could be used to improve the technical quality of Signal School course tests. Test content issues were not addressed because the Signal School closely follows the Systems Approach to Training (SAT) guidelines in order to insure test content, and because Subject Matter Experts indicate that test content is not problematic within the Signal School.

Criterion Referenced Testing at the Signal School

At the Signal School, the development of course tests is similar to the development of Skill Qualification Tests (SQTs). Both types of tests are criterion referenced and are based on task lists. A correspondence between the task lists and the test items is required for both types of tests to insure that the tests are representative of the tasks for that Military Occupational Specialty (MOS). In fact, SMEs at the Signal School are often tasked to revise Signal Programs of Instruction (POIs)

and Signal SQTs, i.e. SMEs are given dual responsibilities. The Signal School has promoted test quality by requiring course test developers to follow SAT guidance and by sponsoring test construction workshops for SQT developers. The primary content difference between the two types of tests is that the SQTs are designed to test a broader range of skills because the field experience of soldiers includes activities that can not be taught or tested at the Signal School.

An important procedural difference between the development and refinement of SQTs and that of course tests is that test item information is provided to SQT developers by the Army Training Support Command (ATSC). Signal School SQT developers use the information to help identify and correct problematic test items. In contrast, test item information is not available for Signal School course test developers and has not been integrated into the SAT guidance on course test development. The successful use of item information by SQT developers suggests that this information might be used to refine Signal course tests.

Description of SQT Test Item Development and Standards

The initial development of SQT items is based on the comparison of the test performance of groups of soldiers that can perform a task, versus groups that can not perform the task. To be included on an SQT, the items must be answered correctly by over 50 percent of the performers and the performers must score higher than the non-performers (M. Andriliunas, personal communication, March 1990, ATSC, Fort Eustis, VA; TRADOC Reg 351-2). This approach is problematic because practical constraints on item development resources limits the size of the two groups to a maximum of ten individuals. It is noteworthy that if this procedure were modified to use large groups, it would insure item consistency by identifying test items that discriminate between groups of soldiers.

After the SQTs have been formally administered to large groups of the soldiers, the ATSC calculates test item statistics. The statistics are returned to each Military Occupational Specialty proponent and are provided to SQT developers for use in the revision of the tests. The ATSC sets standards for the proponents to follow to insure that the MOS proponents utilize similar guidelines during SQT revision.

Although the ATSC recommends reviewing test items and test item distractors on the basis of the item statistics, the ATSC does not require that the test items be changed on the basis of the item statistics. The item statistics and recommendations are designed as an aid to assist SQT developers in identifying problematic test items.

The ATSC uses a computer program to identify questionable test items by monitoring item difficulty, item consistency, and distractor attractiveness. The following paragraphs describe the item statistic standards that have been set for SQT scores.

Item difficulty is defined as the proportion of examinees who correctly answer each test item. The item difficulty value ranges from 0 to 100 percent, indicating that between 0 and 100 percent of the responses were correct for that test item. According to TRADOC Regulation 351-2, test item difficulty should vary between 50 percent and 95 percent. Item difficulty values that are less than 50 percent usually indicate an error in the answer key and are relatively rare. More commonly, an unchallenging or poorly written question will be correctly answered by a high proportion of the examinees. Item difficulty is monitored to insure that the test items measure variance in the content areas that are tested.

Item consistency is estimated by the point biserial correlation between performance on each test item and performance on the remainder of the test. The point biserial will range from -1.0 to 1.0. Positive values indicate that performance on that item is consistent with performance on the rest of the test while negative values indicate that the better test performers were below average for that test item. The item consistency index can be used to identify test questions that discriminate between the better and poorer students. ATSC has the goal of obtaining an item consistency index greater than 0.20 for SQT items (M. Andriliunas, personal communication, March 1990, ATSC, Fort Eustis, VA).

Distractor attractiveness quantifies the extent to which examinees find incorrect distractors plausible on multiple choice tests. This is important because examinees can eliminate implausible distractors and choose the correct answer without adequate knowledge of the content area of the item. According to TRADOC Regulation 351-2, the attractiveness of each distractor should exceed 5 percent. This implies that the item difficulty of multiple choice test items should be less than .85 because the three distractors on a standard four-choice multiple choice test question should be chosen by over 15 percent of the population; however, the less restrictive standard, .95, was adopted to be consistent with the ATSC Item Difficulty standard.

Across Army SQTs, a high proportion of the test items do not meet the ATSC standards; for example, 4,000 of 17,000 recently analyzed test items have item consistencies less than 0.20 while 800 of the items were less than 0.00 (M. Andriliunas, personal communication, March 1990, ATSC, Fort Eustis, VA)

Test Analysis Program Recommendation

In order to provide course developers with item statistics, ARI recommended utilizing the Test Analysis Program (TAP) to analyze the course tests of two Basic Non-commissioned Officer Courses (BNCOCs). The TAP runs on a microcomputer and has been integrated with a Scantron sheet reader. This system allows student responses to be read by a Scantron form reader and placed into a data file. The data are then analyzed by the TAP and item statistics are calculated for each test item. This system can quickly provide the course developer with information that would otherwise be impractical to calculate.

The TAP statistics include estimates of item consistency, item difficulty, and distractor attractiveness. These statistics are very similar to those calculated by ATSC. The main difference between the ATSC system and the TAP is that TAP is designed for micro-computer and can be used at the Signal School by course developers. The TAP can also be used to compute item consistencies for competency areas, thus it is possible to determine if responses on items are consistent with either the entire test, or with related groups of items. The TAP also has item banking capabilities that are designed to produce equivalent forms of tests. This capability could be used by the Signal School to fulfill TRADOC regulations requiring three test versions for each course.

In addition to the item statistics, the TAP computes two single form estimates of test reliability. (Test reliability estimates have implications for power analysis, see Appendix 1.) The reliability estimates are based on the Coefficient Alpha and the Spearman Brown Split-half formulae. The TAP will also compute reliability estimates for test subscales or competency areas.

Method

Two BNCOCs, the 29E and the 31M, were chosen for this evaluation because these courses utilize multiple choice tests and have a higher throughput than other Signal School BNCOCs. The Signal School estimated that 120 students would be trained in the 29E BNCOC and 420 would be trained in the 31M BNCOC during 1989. The multiple choice tests, which these courses use, were adapted for computerized scoring and grading using Scantron forms as answer sheets.

During the data collection period, a total of 212 students were tested in the 31M course while 72 were tested in the 29E course. Each item statistic is based on the test results of a subset of the two sets of students because the Signal School is required to maintain three versions of each annex test to avoid compromising tests and to allow the retesting of students who fail course tests. Only test items, for which more than 20 soldiers were tested, were analyzed for this report.

Results and Discussion

Feedback to the Course Developers

The Test Analysis Program was used to compute item statistics and analyze the responses to the multiple choice tests. For each test item, summary statistics were calculated and were provided to the course developers. Questionable items were flagged in feedback given to the course developers.

The test items were evaluated with standards that are very similar to those used by the ATSC. An item was slated to be flagged if the point biserial correlation was less than 0.20, or if the item difficulty was higher than .95. In addition, item distractors were to be flagged if less than 5 percent of the soldiers chose that distractor.

It was necessary to modify the standards for feedback to the course developers because the original standards led to an extremely high proportion of flagged items. The standards were changed so that no more than half of the items for any one test were flagged. Distractor attractiveness information was provided to the course test developers, but questionable distractors were not individually flagged because most of the test items contained distractors that did not meet the ATSC standard.

Item Statistics Summary

Table 1 summarizes the item statistics by class and contains class estimates of item consistency, item difficulty, and distractor attractiveness for the BNCOC tests. The three distractor attractiveness columns contain the proportion of students who chose the most attractive, second most attractive and the least attractive distractor for each test item. For comparison purposes, Table 1 also summarizes the item statistics for the 29E Skill Level 2 SQTs for 1989.

Table 1. Summary Statistics for the 29E and 31M Tests

Test	Mean Item Diff	Mean Item Cons	Test Size	Sample Size	Percentage Choosing Corr Answ	Distractor 1	Distractor 2	Distractor 3	Reliabt Coeffct Alpha
				29-E					
Av3	90.9	.30	20	46.9	91	7	2	0	.55
Bv2	88.3	.30	30	24.0	88	10	2	0	.54
Bv3	88.0	.23	30	28.9	88	9	2	0	.57
Cv3	92.6	.24	20	42.9	93	6	1	0	.35
E1v2	92.6	.19	40	33.8	93	6	1	0	.55
E1v3	86.0	.31	20	31.0	87	10	2	1	.59
SQT	84.5	.26	118	227.0	84	10	4	1	.80
				31-M					
Av1	85.6	.21	50	210.9	86	11	2	1	.60
D	89.5	.26	50	163.8	90	9	1	0	.75
Fv1	89.3	.14	50	74.8	89	8	2	0	.38
Fv2	86.4	.18	50	34.0	86	10	3	1	.62
Fv3	87.9	.17	50	61.9	88	9	2	0	.42

Table 2 contains the proportion of items that did not meet the ATSC standards for each of three statistics: item difficulty, item consistency, and distractor attractiveness. The table demonstrates that a very high proportion of the items did not meet the standards for each test. This was true regardless of whether the original standards were used to identify questionable test items or whether the standards for the test items were lowered to limit the number of flagged items.

According to Table 2, many of the test items are extremely easy; approximately 37 percent of the course test items have an ease index that is greater than .95 while 72 percent of the items have an ease index greater than .85. Table 2 also indicates that 48 percent of the test items across the course tests have low item consistency estimates, i.e. less than .20.

The distractor attractiveness columns in Table 2 indicate that only 50 percent of the test items have at least one distractor that attracts more than 5 percent of the response. Table 2 also indicates that very few items have more than one attractive distractor as shown by the fact that 89 and 98 percent of the test items do not have second and third distractors that are chosen by more than 5 percent of the students.

A comparison of the item difficulty and item consistency estimates indicates that 78 percent of the items, which do not meet the item difficulty standard (.95), do not meet the item consistency standard. The overlap indicates that the item ease index is nearly as effective as the item consistency measure in identifying items with low consistency estimates. This may be

relevant to test redesign because most SMEs find item difficulty estimates easier to understand and compute than item consistency estimates.

Table 2. Proportion of test items not meeting ATSC Standards

Test	Item Difficulty	Item Consistency	Answer Distribution			
			Correct Response	Distractor 1	2	3
		31M				
Av1	.24	.45	.63	.39	.84	.85
D	.30	.34	.78	.38	.96	1.00
Fv1	.38	.67	.77	.65	.88	1.00
Fv2	.35	.59	.63	.46	.74	.96
Fv3	.33	.64	.71	.53	.89	.98
		29E				
Av3	.47	.29	.76	.53	.94	1.00
Bv2	.52	.52	.57	.52	.87	1.00
B3	.40	.40	.70	.40	.80	1.00
Cv3	.40	.40	.85	.65	1.00	1.00
E1v2	.53	.57	.84	.63	.95	1.00
E3v3	.21	.37	.63	.37	.89	1.00
		Summary				
Mean CTs	.37	.48	.72	.50	.89	.98
29E-SQT	.20	.33	.56	.31	.74	.96
ATSC Standards	50%-95%	>.20	<85%	>5%	>5%	>5%

The relationship between item difficulty and item consistency also indicates that the more difficult questions are more consistent with overall test performance. This suggests that while the procedures followed by the course developers are adequate to insure item consistency, too many of the items are not sufficiently challenging. By increasing the difficulty of the test items, it can be expected that the item consistency estimates will also increase.

The distractor attractiveness data are relevant to the issue of increasing item difficulty because the data suggest that many Signal test items utilize distractors that are not plausible to the users. In effect, many of the test items function as two choice rather than four choice questions because two of the distractors are not reasonable choices. It follows that improving course test item distractor attractiveness would produce more challenging and useful questions.

Tables 1 and 2 allow the comparison of test item statistics obtained for the SQTs and the course tests. The tables indicate that SQTs have more acceptable item characteristics than the

course tests. The difference in item statistics is also reflected by the higher reliability estimates of the SQTs. Given the similarity in expertise between the SQT developers and the course test developers, a major reason for this difference may be the availability of item information to the SQT developers. This interpretation is confirmed through informal feedback with course developers, who report that the item information is useful in identifying poor test items.

Conclusions

Data from the TAP were utilized to identify questionable items and item distractors. By using the TAP, it can be expected that better test item distractors will be identified and that more challenging test items will be created. This process will help insure that Signal School course tests can be used for the validation of lessons and courses. This conclusion is consistent with the comparisons between the technical quality of the SQTs and the course tests and is reinforced by reports from Signal School SMEs that the item information is useful in identifying questionable test items.

The comparisons between the technical qualities of the SQTs and the course tests indicate that test items could be improved by providing distractor attractiveness and item difficulty information to Signal School course test developers. The impact of providing item consistency data would probably be minimal because most of the items that are flagged for item consistency are also flagged for item difficulty and distractor attractiveness. The analyses conform to the view of Signal School SMEs that the content of the test items is adequate, and that the test items should be designed to be more challenging.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. Educational Researcher, 13, 4-16.

Cohen, J. (1977). Statistical power analyses for the behavioral sciences. New York, NY: Academic Press.

Gulliksen, H. O. (1950). Theory of mental tests. New York, NY: Wiley.

Jensen, A. R. (1980). Bias in mental testing. New York, NY: Free Press.

Lord, F. & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison Wesley.

McNemar, Q. (1969). Psychological statistics. New York, NY: Wiley.

TRADOC REGULATION 350-7. (1988). Training: Systems approach to training.

TRADOC REGULATION 351-2. (1986). Schools: Skill qualification test and common task test development policy and procedures.

Appendix A.

Relationship Between Test Reliability and Training Evaluation

Test reliability is of limited importance if the test is used solely for the purpose of determining whether the performance of a student has reached an agreed-upon criterion. However, the Signal School has recognized the importance of using the tests for other purposes, such as training evaluation. Test reliability limits the ability of the researcher to evaluate the effectiveness of new training procedures by making group differences harder to demonstrate and by underestimating the magnitude of an experimentally induced effect.

One effect of a deficiency in test reliability is that the sample size needed to evaluate a new training procedure will increase. The increase occurs because an inference can only be concluded when the ratio of mean differences between groups to mean score variance exceeds some constant. For example the t-test is given by McNemar (1969) as:

$$t\text{-ratio}=(Mean_{y1}-Mean_{y2})/s_{mean\text{-}y}, \text{ where } s_{mean\text{-}y}=s_y/SQRT(N).$$

Note that a decrease in reliability is equivalent to an increase in observed score variance as shown by (e.g. Gulliksen, 1950):

$$r = sd_t^2/sd_x^2,$$

where r equals the test's reliability, sd_x measures the test's variance, and sd_t estimates the variance that would have been obtained had a perfectly reliable test been used.

It follows from the definition of $s_{mean\text{-}y}$ that a decrease in test reliability may be offset by an increase in sample size, N. Thus the data collection cost of a training evaluation will increase when low reliability tests are used.

A second effect of low test reliability occurs because the effect size of a new training procedure relative to an alternative approach is calculated as the ratio of the mean difference between the groups to the square root of variance of the test (Cohen, 1977; Bloom 1984):

$$ES = (m1-m2)/sd_x.$$

The relationship between test reliability and variance is given by (e.g. Gulliksen, 1950):

$$r = sd_t^2/sd_x^2,$$

where r equals the test's reliability, sd_x measures the test's variance, and sd_t estimates the variance that would have been obtained had a perfectly reliable test been used. Because a decrease in test score reliability results in an increase in

test score variance, i.e. sd_x, it follows that attenuation of the reliability of a test leads to an underestimation of the effect size that is being calculated.

It is noteworthy that these formulae can be rearranged to correct the effect size estimate for attenuation of reliability:

$$ES_{corrected} = (m1-m2)/sd_t = (m1-m2)/(sd_x*SQRT(r)).$$

Obviously, this estimate is most credible when a liberal estimate of test reliability is used.

The Test Analysis Program uses coefficient alpha and the split-half approach to estimate a test's reliability. Both of these estimates are frequently used to estimate test reliability because they represent a lower bound on test reliability (Jensen, 1980).

An alternate approach to estimating test reliability is to separately calculate coefficient alpha for each subscale or content area. The subscale reliability estimates and content correlations may then be used to obtain a higher estimate of the test's reliability. This formula is most useful when estimating the reliability of a heterogenous test because the effect of low subscale correlations is minimized. The approach and methodology can be found in Lord and Novick (1974).